Adaptive importance sampling by kernel smoothing

Bernard Delyon and François Portier*

Université de Rennes 1 and Institut Polytechnique de Paris

March 22, 2019

Abstract

A key determinant of the success of Monte Carlo simulation is the *sampling policy*, the sequence of distribution used to generate the particles, and allowing the sampling policy to evolve adaptively during the algorithm provides considerable improvement in practice. The issues related to the adaptive choice of the sampling policy are addressed from a functional estimation point of view. The considered approach consists of modelling the sampling policy as a mixture distribution between a flexible *kernel density estimate*, based on the whole set of available particles, and a naive heavy tail density. When the share of samples generated according to the naive density goes to zero but not too quickly, two results are established. Uniform convergence rates are derived for the sampling policy estimate. A central limit theorem is obtained for the resulting integral estimates. The fact that the asymptotic variance is the same as the variance of an "oracle" procedure, in which the sampling policy is chosen as the optimal one, illustrates the benefits of the proposed approach.

Keywords: Monte Carlo methods; adaptive importance sampling; kernel density estimation; uniform convergence rates; martingale methods.

1 Introduction

Computing integrals is an important issue taking place in many domains of science and the Monte Carlo simulation framework has became indisputably useful especially when the underlying dimension is large. The general issue tackled in this paper is that of integral approximation. More specifically, quantities of interest are of the form $\int gf$, for possibly many integrands g and a single positive probability density function $f : \mathbb{R}^d \to \mathbb{R}_{>0}$ (the underlying measure is the Lebesgue measure). The approach taken, *adaptive importance sampling* (AIS), is made of the following ingredients:

(i) The particles $(X_i)_{i=1,\ldots,n}$ are chosen at random according to the *sampling policy*, denoted by $(q_i)_{i=0,\ldots,n-1}$, a collection of probability density functions, in that each X_i is distributed according to q_{i-1}. The resulting AIS integral estimate of $\int gf$ takes the form $\sum_{i=1}^n w_{n,i} g(X_i)$ with $w_{n,i} \propto f(X_i)/q_{i-1}(X_i)$ such that $\sum_{i=1}^n w_{n,i} = 1$.

(ii) The sampling policy $(q_i)_{i=0,\ldots,n-1}$, which (usually) estimates f (Oh and Berger, 1992, section 1.2), will evolve sequentially during the algorithm in order to take into account the new pieces of information obtained from the last calls to f. A crucial point is the ability of the algorithm to estimate f rapidly. This is the "adaptive" character of the method.

*Corresponding author, francois.portier@gmail.com.

Classically, two families of methods can be distinguished depending on the approach taken to model the sampling policy: *parametric* and *nonparametric*. When $(q_i)_{i=0,\dots,n-1}$ is a constant sequence, it corresponds to *independent importance sampling*. See for instance (Evans and Swartz, 2000, Chapter 6) or (Robert and Casella, 2004, chapter 3).

(iii) From an algorithmic point of view, it may be useful to freeze the sequence q_i along some "blocks". This will save the time needed to update and allow to run in parallel the generation of the random variables according to q_i.

Pioneer works on adaptive schemes have focused on parametric families to model the sampling policy. They include, among others Kloek and Van Dijk (1978), Geweke (1989), Oh and Berger (1992), Owen and Zhou (2000), Cappé et al. (2004), Cappé et al. (2008) (see also Elvira et al. (2015) for a review on the variant called *adaptive multiple importance sampling*). In Oh and Berger (1992), martingale techniques were successfully employed to describe AIS schemes and their approach was recently extended (Delyon and Portier, 2018) to obtain a central limit theorem for AIS integral estimates when q_i is chosen out of a parametric family. They established that the asymptotic variance is the same as the asymptotic variance of an oracle strategy in which $q_i = f$ for all $i = 0, \dots, n-1$. First research works to consider more flexible nonparametric approaches than parametric AIS were based on kernel smoothing techniques and include West (1993), Givens and Raftery (1996), Zhang (1996) (see also Neddermeyer (2009) about the algorithmic efficiency). These authors investigated a similar approach in that they use a kernel density estimate based on the previous particles re-weighted by importance weights. In particular, Zhang (1996) studies (non-normalized) AIS integral estimates of $\int gf$ for a single positive function g and when the sampling policy q_i targets the density proportional to gf (rather than f as described before).

The asymptotic analysis of AIS is classically executed according to the total amount of generations n going to ∞. Given that, a first asymptotic regime is when the sequence $(q_i)_{i\geqslant 0}$ is frozen from a certain given time, i.e., the number of updates is finite. In this regime, central limit theorems are given in Chopin (2004), Douc et al. (2007a,b). In Zhang (1996), the author considers nonparametric AIS and works under a second asymptotic regime in which the number of samples coming from the first sampler goes to ∞ so that it represents a fixed share among the whole sample. A third asymptotic regime, which better suits the true nature of these adaptive schemes is when the algorithm never stops updating, i.e., the number of updates goes to infinity. In this context, the consistency of parametric adaptive multiple importance sampling (a variant of AIS) has been obtained recently in Marin et al. (2019). They require that the number of samples between each update, i.e., the size of the blocks defined in (iii), is an increasing sequence going to ∞. For parametric AIS, the central limit theorem in Delyon and Portier (2018) is obtained without any restriction on the size of the blocks. Note that some different uniform consistency results are presented in Feng et al. (2018).

The novelty of the paper is to consider the issue of *flexible* sampling policy learning as a functional approximation problem and the main contribution is to establish uniform convergence rates estimating f with possibly unbounded support and without any condition on the blocks. The obtained convergence rates are sharp in the sense that they are the same as the one obtained in the problem of nonparametric density estimation with independent random variables (Giné and Guillou, 2001), that is $\sqrt{\log(n)/(nh_n^d)}$, up to a bias term, where h_n is the bandwidth parameter. The proposed sampling policy estimate is a mixture between a kernel smoothing estimate of f (similar to West (1993), Givens and Raftery (1996), Zhang (1996), Neddermeyer (2009)) and some fixed naive density with heavy tails compared to the ones of f. The tuning of the mixture parameter between this two densities will be a key ingredient to obtain the stated rate of convergence. The mixture parameter underlines a classical trade-off (Owen and Zhou, 2000, section 2.3) between the variance efficiency, that is achieved when q_i is close to f (as detailed in the next section), and the exhaustiveness of the visit, that is achieved when q_i has sufficiently large tails.

From a theoretical point of view, the critical aspect of this work is to deal with kernel smoothing estimates in an adaptive environment characterized by a particular dependence structure of the random variables of interest. The proposed approach bears ressemblance with the one developped in Delyon and Portier (2018) where martingale tools have been used to study parametric AIS but more powerful results are required. Specifically, to handle kernel based estimates with importance weights, a modified version of Bennett's concentration inequality (Freedman, 1975) turns out to be very useful. Our results are then related with the ones dealing with kernel density estimates for independent sequences (Giné and Guillou, 2001, 2002), for weak dependent sequences (Hansen, 2008), and for Markov chains (Azaïs et al., 2018; Bertail and Portier, 2018). In these papers, the same rate of convergence, $\sqrt{\log(n)/(nh_n^d)}$ (for the variance term), was obtained but under different assumptions on the dependence structure of the considered sequences.

Another contribution of the paper is a central limit theorem for the resulting integral estimate. As in Delyon and Portier (2018), the asymptotic variance is the same as the variance of the "oracle" procedure that would use $q_i = f$ from the beginning.

The outline is as follows. In section 2, the proposed algorithm is presented and illustrated. The martingale tools that will be the basis of our analysis are introduced in section 3. The main results are stated in section 4 and some concluding comments are given in section 5. An appendix section is devoted to the mathematical proofs.

2 The algorithm

Let $f : \mathbb{R}^d \to \mathbb{R}_{>0}$ be a positive probability density function. Let $(q_i)_{i=0,\ldots,n-1}$ be a collection of probability density functions and $(X_i)_{i=1,\ldots,n}$ be a collection of random variables generated according to the sampling policy $(q_i)_{i=0,\ldots,n-1}$, i.e., each $X_i \sim q_{i-1}$. The proposed approach is to estimate f by a mixture of $n \geqslant 1$ densities each having mean X_i and standard deviation h_i. Specifically, each component $i \in \{1, \ldots, n\}$ of the mixture is given by $x \mapsto K_{h_i}(X_i - x)$ where $K_h(u) = K(u/h)/h^d$, $K : \mathbb{R}^d \to \mathbb{R}_{\geqslant 0}$ is a density called *kernel* and h_i is a positive number called *bandwidth*. The kernel estimate of f at step $n \geqslant 1$, f_n, is given by

$$f_n(x) = \sum_{i=1}^{n} w_{n,i} K((x - X_i)/h_i)/h_i^d, \qquad x \in \mathbb{R}^d, \tag{1}$$

where $(w_{n,i})_{i=1,\ldots n}$ is a vector of *importance weights* defined by

$$\forall i = 1, \ldots, n, \qquad w_{n,i} \propto w_i = \frac{f(X_i)}{q_{i-1}(X_i)} \quad \text{such that} \quad \sum_{i=1}^{n} w_{n,i} = 1.$$

Each weight reflects the importance of the associated particle within the mixture. Note in passing that, given a collection $(q_i)_{i=0,\ldots,n-1}$, evaluating f_n only requires to know f up to a scale factor. This kind of estimates is called *self-normalized importance sampling* (Owen, 2013, Chapter 9).

Any sampling policy $(q_i)_{i=0,\ldots,n-1}$ defines an AIS algorithm but the most relevant ones should satisfy these two requirements:

(a) **unbiasedness**. The sampling policy $(q_i)_{i=0,\ldots,n-1}$ is said unbiased if for all $i = 1, \ldots, n$ and for all bounded measurable function $g : \mathbb{R}^d \to \mathbb{R}_{\geqslant 0}$, it holds that $\mathbb{E}[f(X_i)g(X_i)/q_{i-1}(X_i)] =$

$\int gf$. When f does not vanish, this is actually satisfied if and only if each q_i is supported on the whole space $\mathbb{R}^d$ (when f is *not* supported on the whole space $\mathbb{R}^d$, see Evans and Swartz (2000), Section 6.2). This condition on the support allows to visit the whole space of interest.

(b) **minimum variance**. Define

$$V(q,g) = \int \frac{g^2 f^2}{q} - \left(\int gf \right)^2 . \tag{2}$$

The quantity $V(q,g)$ corresponds to a fixed-policy variance in that $nV(q,g)$ is the variance of $\sum_{i=1}^n g(X_i)f(X_i)/q(X_i)$ when the $(X_i)_{i=1,\dots,n}$ are independently and identically distributed according to q. Hence it is in our interest to choose q achieving the minimum of the following integrated-variance criterion

$$C(q) = \int_{\mathbb{R}^d} V\Big(q, K\big((x-\cdot)/h\big)\Big)\, \mathrm{d}x.$$

Fortunately, this minimum is unique and is achieved when $q = f$ (whatever the value of h) as stated in the following lemma, whose proof is given in the Appendix

Lemma 1. *The minimum of C over the positive probability densities on $\mathbb{R}^d$ is achieved if and only if the sampler is $q = f$.*

This result bears resemblance with Theorem 6.5 in Evans and Swartz (2000) that provides the optimal sampler when integrating a single function. Hence the sampling policy should be similar to f. This differs from Zhang (1996) where the targeted policy is not f and depends on the integrand.

Balancing between both conditions, (a) and (b), our proposal for the sampling policy $(q_i)_{i=0,\dots,n-1}$ is

$$q_i(x) = (1-\lambda_i) f_i(x) + \lambda_i q_0(x), \qquad x \in \mathbb{R}^d, \tag{3}$$

where $(\lambda_i)_{i=0,\dots,n-1} \subset [0,1]$ is a sequence of mixture weights. Initialization is given by $\lambda_0 = 1$, $f_0 = 0$, and q_0, the initial sampler, is supposed to be supported over $\mathbb{R}^d$. In the mixture, the component q_0 permits to visit the space extensively during the algorithm, ensuring condition (a) as soon as q_0 has a large support. On the other side the value of λ_i shall decrease during the procedure in order to gain in efficiency, as condition (b) indicates. Balancing suitably between f_i and q_0 permits to realize the trade-off, described in (Owen and Zhou, 2000, section 2.3), between tentatively optimal and defensive strategy. Note in passing that generating from K and q_0 allows to generate according to q_i. The algorithm is written below and an illustration is provided in Figure 1.

Algorithm.
Inputs: *The bandwidths $(h_i)_{i=1,\dots,n}$, the mixture weights $(\lambda_i)_{i=0,\dots,n-1}$, the initial density q_0.*

For $i = 1, 2, \dots, n$:

generate X_i from q_{i-1} defined in (3) and compute $w_i = f(X_i)/q_{i-1}(X_i)$

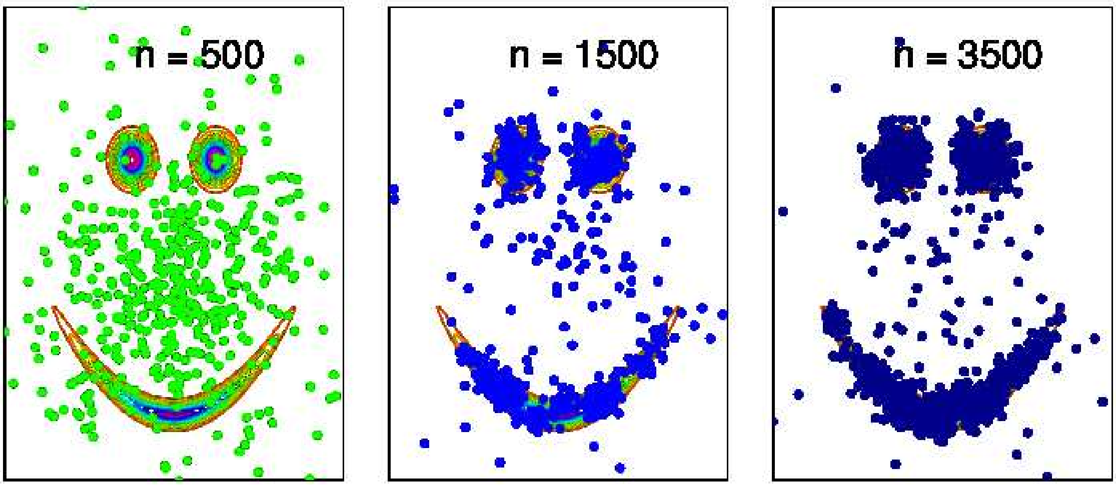

Figure 1: Sampling policy learning based on 3500 particles when the targeted policy f is defined on $\mathbb{R}^2$ as a mixture of a banana shape distribution and two Gaussians. The mixture weights λ_i are equal to 1 for $i = 0, \ldots, 499$, to 0.05 for $i = 500, \ldots, 1499$ and to 0.01 for $i = 1500, \ldots, 3499$. The bandwidths h_i are equal to 1 all along the algorithm. The initial density q_0 is a standardized Student's t-distribution. The particles indexed by $i = 1, \ldots, 500$ are plotted on the left, $i = 501, \ldots, 1500$ are plotted in the middle, $i = 1501, \ldots, 3500$ are plotted on the right.

As discussed in the introduction, a usual extra ingredient of the proposed algorithm is to divide the n particles into "blocks" of particles having the same distribution. In other words, the sampling policy q_i is frozen over these blocks and the update of q_i is conducted only when i hits some update set. In the rest of the paper, for clarity reasons, we only consider the algorithm described before that updates the sampling policy at each sample. The extension to arbitrary update set can be carried out by modifying our proofs.

The *sequential Monte Carlo* approach (Chopin, 2004; Del Moral et al., 2006) is based on generating new particles around each existing ones. While in view of (1), it is similar in spirit to our approach of sequential kernel estimation, it is not included in our framework simply because the distribution of the particles does not take the form of (3).

3 The martingale framework

Given the definition of f_n in (1), quantities of interest are of the form

$$\sum_{i=1}^{n} w_{n,i} g(X_i) = \frac{I_n(g)}{I_n(1)} \qquad \text{with} \qquad I_n(g) = n^{-1} \sum_{i=1}^{n} w_i g(X_i)$$

where $g : \mathbb{R}^d \to \mathbb{R}$. By virtue of the decomposition

$$\left(\sum_{i=1}^{n} w_{n,i} g(X_i) - \int gf \right) = \frac{1}{I_n(1)} \left((I_n(g) - \int gf) - (I_n(1) - 1) \int gf \right), \tag{4}$$

as soon as we handle terms of the form $\sum_{i=1}^{n} w_i g(X_i)$ classical preservation tools from asymptotic theory (van der Vaart, 1998) will ensure to control $\sum_{i=1}^{n} w_{n,i} g(X_i)$. For this reason, a key quantity in the following is

$$M_n(g) = \sum_{i=1}^{n} \left\{ w_i g(X_i) - \int gf \right\}.$$

The starting point of our approach is the martingale property verified by $M_n(g)$ associated with the natural σ-field $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, $\mathcal{F}_0 = \emptyset$.

Lemma 2. *If for each $i \geqslant 1$, q_i is supported on $\mathbb{R}^d$ and $\int |g|f < \infty$, the sequence $(M_n(g), \mathcal{F}_n)$ is a martingale with quadratic variation $\langle M_n(g) \rangle = \sum_{i=1}^{n} V(q_{i-1}, g)$.*

We now recall two classical results from martingales theory. The first one is a concentration inequality for martingale arrays. It is a modification of (Freedman, 1975, Theorem 4.1), allowing the martingale increments to be unbounded. Because this result takes into account the rate of decrease of the quadratic variation (v appears in the bound), it plays a crucial role to control the behavior of kernel estimator (in proving Theorem 6) for which the quadratic variation will depend on the bandwidth h_n.

Theorem 3. *Let $(Y_i)_{1\leqslant i \leqslant n}$ be random variables such that*

$$\mathbb{E}[Y_i \mid \mathscr{F}_{i-1}] = 0, \quad \text{for all } 1 \leqslant i \leqslant n,$$

then, for all $t \geqslant 0$ and $v, m > 0$,

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} Y_i \right| \geqslant t,\ \max_{i=1,\ldots,n} |Y_i| \leqslant m,\ \sum_{i=1}^{n} \mathbb{E}[Y_i^2 \mid \mathscr{F}_{i-1}] \leqslant v \right) \leqslant 2 \exp\left(-\frac{t^2}{2(v + tm/3)} \right).$$

The following result is central limit theorem for martingale array. It will be useful to prove our central limit theorem, Theorem 7.

Theorem 4. *(Hall and Heyde, 1980, Corollary 3.1) Let $(W_{n,i})_{1\leqslant i \leqslant n,\, n \geqslant 1}$ be a triangular array of random variables such that*

$$\mathbb{E}[W_{n,i} \mid \mathscr{F}_{i-1}] = 0, \quad \text{for all } 1 \leqslant i \leqslant n, \tag{5}$$

$$\sum_{i=1}^{n} \mathbb{E}[W_{n,i}^2 \mid \mathscr{F}_{i-1}] \to v^* \geqslant 0, \quad \text{in probability}, \tag{6}$$

$$\sum_{i=1}^{n} \mathbb{E}[W_{n,i}^2 \mathrm{I}_{\{|W_{n,i}| > \varepsilon\}} \mid \mathscr{F}_{i-1}] \to 0, \quad \text{in probability}, \tag{7}$$

then, $\sum_{i=1}^{n} W_{n,i} \rightsquigarrow \mathcal{N}(0, v^*)$.

4 Main results

The results of the paper are expressed using the sequences

$$a_n = \sqrt{\frac{\log(n)}{nh_n^d}} \qquad \text{and} \qquad \overline{h_n^2} = n^{-1} \sum_{i=1}^{n} h_i^2.$$

The approach taken to obtain the uniform convergence rate $\sqrt{\log(n)/(nh_n^d)}$ for f_n proceeds in 2 steps. A first step (Lemma 5) is required to derive initial bounds on the behavior of f_n. These bounds will then permit to have an accurate control of $w_i = f(X_i)/q_{i-1}(X_i)$, appearing in the sums, and allow to finally obtain the stated convergence rate in $a_n + \overline{h_n^2}$. These initial bounds, which are interesting in their own, are stated in the following lemma. They are valid under weaker conditions, given below, than the ones needed to establish the final result.

(H1) $f : \mathbb{R}^d \to \mathbb{R}_{\geqslant 0}$ is a probability density function. The function f is bounded by U_f and f/q_0 is bounded by U_{f,q_0}.

(H2) $K : \mathbb{R}^d \to \mathbb{R}_{\geqslant 0}$ is a probability density function bounded by U_K. In addition,

$$\int uK(u)\,\mathrm{d}u = 0, \qquad \int u^2 K(u)\,\mathrm{d}u + \int K(u)^2\,\mathrm{d}u < \infty,$$

and there exists $C_K > 0$ such that for any $0 < \eta < 1$,

$$\int_{\mathbb{R}^d} \sup_{0 \leqslant \|u\| \leqslant \eta} |K(y) - K(y+u)|\,\mathrm{d}y \leqslant C_K \eta.$$

(H3) The function f is two times continuously differentiable, with bounded second derivatives.

Lemma 5 (initial bound)**.** *Under (H1), (H2) and (H3), if λ_n and h_n are positive decreasing sequences such that $a_n^2 \ll \lambda_n$, then we have, for any $r > 0$,*

$$\sup_{n \geqslant 1} \sup_{\|y\| \leqslant n^r} \frac{|f_n(y) - f(y)|}{a_n \lambda_n^{-1/2} + \overline{h_n^2}} < \infty, \qquad a.s.$$

The convergence rate $a_n \lambda_n^{-1/2} + \overline{h_n^2}$ obtained in the previous lemma is too large compared to $a_n + \overline{h_n^2}$ whenever λ_n goes to 0. To obtain the appropriate rate of convergence we need some additional regularity conditions that express the relationship needed between f and q_0.

(H4) For any $\varepsilon \leqslant 1$, $f(x)^\varepsilon / q_0(x)$ is bounded by U_{f^ε, q_0}. There exists (k_0, C_0), positive numbers, such that $q_0(x) \leqslant C_0(1 + \|x\|^{-k_0})$.

Theorem 6 (Uniform convergence rate)**.** *Under (H1), (H2), (H3) and (H4), if λ_n and h_n are positive decreasing sequences such that there exists $\delta > 0$ for which $a_n^{1-\delta} \ll \lambda_n$, $\overline{h_n^2}^{1-\delta} \ll \lambda_n$, then, as $n \to \infty$,*

$$\sup_{x \in \mathbb{R}^d} |f_n(x) - f(x)| = O_{\mathbb{P}}\left(a_n + \overline{h_n^2}\right).$$

Let us now state a central limit theorem for the integral estimates $\sum_{i=1}^n w_{n,i} g(X_i)$. The function g needs to satisfy the following assumption.

(H5) There exists $k \geqslant 1$, $\int g^2 q_0^k < \infty$.

Theorem 7 (asymptotic optimality)**.** *Under (H1), (H2), (H3), (H4) and (H5), if λ_n and h_n are positive decreasing sequences going to 0 such that there exists $\delta > 0$ for which $a_n^{1-\delta} \ll \lambda_n$, $\overline{h_n^2}^{1-\delta} \ll \lambda_n$, as $n \to \infty$,*

$$\sqrt{n}\left(\sum_{i=1}^n w_{n,i} g(X_i) - \int gf\right) \rightsquigarrow \mathcal{N}(0, V(f,g)).$$

5 Concluding remarks

Choice of (λ_n, h_n). Balancing the variance term a_n and the bias term $\overline{h_n^2}$ in Theorem 6 leads to $h_n \simeq n^{-1/(4+d)}$ (up to a logarithm). This corresponds to the usual optimal rate in non parametric estimation when the function is at least 2-times continuously differentiable and the kernel has order 2.

Allowing λ_n to go to 0 does not change the obtained rate of convergence in Theorem 6 but allows to get the right variance, $V(f,g)$, in Theorem 7. If λ_n were constant one would get $V(q,g)$ (eq. (2)) with $q = (1-\lambda)f + \lambda q_0$; this is easily obtained with a slight modification of the proof of Theorem 7. In practice, a slow decrease of λ_n might be appropriate when facing a difficult problem e.g., several modes or large variance of f.

Asymptotic optimality. Allowing λ_n to go to 0 is the greatest difficulty in the proofs of Theorems 6 and 7. It allows in practice to gain in efficiency by not generating too much under the naive distribution q_0. The total amount of exploratory points drawn under q_0 is approximately $\sum_{i=1}^n \lambda_i$ which is negligible before n. Asymptotically, the effect of q_0 disappears as expressed by the variance $V(g,f)$ in Theorem 7. This contrasts with the results of Zhang (1996) in which the initial sampling policy is of dominant importance compared to the others.

Choice of the kernel. The kernel K being non-negative (this is needed to ensure random generation according to q_n), it can't have more than one vanishing moment i.e., $\int uK(u)\,\mathrm{d}u = 0$. This bounds the exploitable smoothness of f to two derivatives and explains why the rate of decrease of the bias term is in $\overline{h_n^2}$.

In Hansen (2008), a condition, stronger than (H2), on the kernel K is used to derive similar results (under mixing assumptions); it is $|K(y) - K(y+u)| \leqslant \|u\| K^*(y)$, where $\int K^*(y)\,\mathrm{d}y < \infty$. Many popular kernels satisfy this condition: Student, Gaussian, Epanechnikov, Quartic, Triangular. Our setting permits to include the classical uniform kernel $K = \mathrm{I}\{[-1/2, 1/2]^d\}$.

Different settings. Condition (H4) is stronger than what would be strictly necessary. In fact, one could ask the function f^ε/q_0 to be bounded (in the tail) for a single value of ε, but this would require a proper choice of the bandwidth sequence and the weights sequence.

When f is compactly supported and bounded away from 0, the study of the algorithm is simpler and the same results are valid under weaker conditions on λ_n and h_n. This is presented in Appendix B.

Updating q_0. An interesting variant of the proposed approach is to allow q_0 to be updated. For instance, it is tempting to use the current average estimate $\mu_n = \sum_{i=1}^n w_{n,i} X_i$ by replacing q_0 by $q_0(\cdot - \mu_n)$.

Weighted schemes. The use of some update set (as described just below the algorithm in section 2) permits to define a class of weighted estimates that might be more efficient in practice (Douc et al., 2007b; Delyon and Portier, 2018; Owen and Zhou, 2019). Let $\widehat{I}_t$ denote the self-normalized importance sampling estimate at step t associated to sampler q_{t-1}. The weighted estimate takes the form $\widehat{I}_T = \sum_{i=1}^T \alpha_t \widehat{I}_t$, where the α_t are to be chosen, based for instance on an estimate of the variance of $\widehat{I}_t$.

Appendix A Proofs of the stated results

A.1 Proof of Lemma 1

Recall that $K_h(u) = K(u/h)/h^d$. For each $x \in \mathbb{R}^d$, we have

$$V(q, K_h(x - \cdot)) = \int \frac{f(y)^2 K_h(x-y)^2}{q(y)} \,\mathrm{d}y - (f * K_h)(x)^2.$$

Integrating over x leads to

$$C(q) = v_K h^{-d} \int \frac{f(y)^2}{q(y)} \,\mathrm{d}y - \int (f * K_h)(x)^2 \,\mathrm{d}yx.$$

with $v_K = \int K(u)^2 du$. Applying Theorem 6.5 in Evans and Swartz (2000), the minimum of $\int f(y)^2/q(y)\,\mathrm{d}y$ is achieved uniquely when $q = f$.

$\square$

A.2 Proof of Theorem 3

Let us recall the Bennett inequality for supermartingales as given by Freedman in (Freedman, 1975, Theorem 4.1):

Theorem 8. *Let $(X_i)_{1 \leqslant i \leqslant n}$ be a sequence of random variables such that*

$$X_i \leqslant 1 \ \text{ a.e.} \quad \textit{and} \quad \mathbb{E}[X_i \mid \mathscr{F}_{i-1}] \leqslant 0 \ \text{ a.e.} \quad \textit{for all } i,$$

then, for all $a \geqslant 0$ and $b > 0$,

$$\mathbb{P}\Big(\sum_{i=1}^n X_i \geqslant a, \sum_{i=1}^n \mathbb{E}[X_i^2 \mid \mathscr{F}_{i-1}] \leqslant b\Big) \leqslant \exp\Big(-bh(a/b)\Big)$$
$$h(u) = (1+u)\log(1+u) - u.$$

Let us recall also the classical inequality allowing to switch from the Bennett inequality to the Bernstein inequality (Boucheron et al. (2013) p.38 or Pollard (1984) p.193):

$$h(u) \geqslant \frac{u^2}{2(1+u/3)}.$$

By the Jensen inequality, the variables $X_i = \min(Y_i/m, 1)$ satisfy the assumptions of Theorem 8 and we get in particular, since $X_i^2 \leqslant Y_i^2/m^2$,

$$\begin{aligned} \mathbb{P}\Big(\sum_{i=1}^n Y_i \geqslant t, \max_{i=1,\dots,n} |Y_i| \leqslant m, \sum_{i=1}^n \mathbb{E}[Y_i^2 \mid \mathscr{F}_{i-1}] \leqslant v\Big) \\ \leqslant \mathbb{P}\Big(\sum_{i=1}^n X_i \geqslant t/m, \sum_{i=1}^n \mathbb{E}[X_i^2 \mid \mathscr{F}_{i-1}] \leqslant v/m^2\Big) \\ \leqslant \exp\left(-\frac{t^2}{2(v+tm/3)}\right). \end{aligned}$$

By the symmetry of the assumptions on (Y_i), the same inequality holds true with $-Y_i$ instead of Y_i and we get the stated bound. $\square$

A.3 Proof of Lemma 5

Define

$$\begin{aligned}
\widetilde{f}_n &= n^{-1} \sum_{i=1}^{n} f \star K_{h_i}, \\
M_n &= \sum_{i=1}^{n} \{w_i - 1\}, \\
Z_n(x) &= \sum_{i=1}^{n} \left\{ w_i K_{h_i}(x - X_i) - \int f(y) K_{h_i}(x-y)\, \mathrm{d}y \right\}.
\end{aligned}$$

The proof will follow from both forthcoming lemmas (whose proofs are postponed to the end of the section).

Lemma 9. *Under (H2) and (H3), it holds that* $\left\| \widetilde{f}_n - f \right\|_\infty = O(\overline{h_n^2})$.

Lemma 10. *Under (H1) and (H2), if λ_n is a positive decreasing sequence such that* $\log(n)/n = O(\lambda_n)$, *we have* $M_n = O(\lambda_n^{-1/2}\sqrt{n \log n})$, *almost surely. If moreover, h_n is a positive decreasing sequence such that* $\log(n)/nh_n^d = O(\lambda_n)$, *we have, for any* $r > 0$, $\sup_{\|x\| \leqslant n^r} |Z_n(x)| = O(\lambda_n^{-1/2}\sqrt{n \log(n)/h_n^d})$, *almost surely.*

The proof of Lemma 5 follows form the classical bias-variance decomposition:

$$f_n(y) - f(y) = (f_n(y) - \widetilde{f}_n(y)) + (\widetilde{f}_n(y) - f(y)). \tag{8}$$

Similar to (4), it holds that

$$f_n(y) - \widetilde{f}_n(y) = \frac{Z_n(y) - \widetilde{f}_n(y) M_n}{M_n + n}. \tag{9}$$

We can apply Lemma 10 (using that $\log(n)/nh_n^d \ll \lambda_n$) to obtain that with probability 1,

$$Y_n = \sup_{\|y\| \leqslant n^r} \frac{|f_n(y) - \widetilde{f}_n(y)|}{a_n \lambda_n^{-1/2}} = O(1).$$

Equivalently, $\sup_{n \geqslant 1} Y_n < \infty$. Combined with Lemma 9, we obtain the desired result. □

Proof of Lemma 9. Write

$$\begin{aligned}
\left|\widetilde{f}_n(x) - f(x)\right| &= \frac{1}{n}\Big| \sum_{i=1}^{n} \int \big(f(x + h_i y) - f(x)\big) K(y) dy \Big| \\
&= \frac{1}{n}\Big| \sum_{i=1}^{n} \int \big(f(x + h_i y) - f(x) - h_i \langle y, \nabla f(x) \rangle \big) K(y) dy \Big| \\
&\leqslant \frac{\|\nabla^2 f(x)\|_\infty}{2n} \sum_{i=1}^{n} h_i^2.
\end{aligned}$$

□

Proof of Lemma 10. The proof of the first result follows from Theorem 3 with $Y_i = f(X_i)/q_{i-1}(X_i) - 1$: Since w.p.1. $|Y_i| \leqslant m = \lambda_n^{-1} U_{f,q_0} + 1$ and the quadratic variation is $\leqslant v = n\lambda_n^{-1} U_{f,q_0}$, we get

$$\mathbb{P}\big(|M_n| \geqslant t\big) \leqslant 2\exp\left(-\frac{Ct^2\lambda_n}{n+t}\right),$$

for some constant C depending on f and q_0. We conclude by taking $t = \gamma\sqrt{n\log n/\lambda_n}$ for γ large enough and use the Borel-Cantelli lemma.

Consider the second statement dealing with Z_n. Let $\varepsilon_n = h_n n^{-1/2}$ and $(x_i)_{i=1,\ldots N}$ be an ε_n-grid over $\{\|x\| \leqslant n^r\}$, i.e., $\min_{k=1,\ldots N} \|x - x_k\| \leqslant \varepsilon_n$ if $\|x\| \leqslant n^r$. Such a grid can be constructed using $N = C_d \varepsilon_n^{-d} n^{rd}$ points, with $C_d > 0$. Define

$$r_n(y) = \sup_{0 \leqslant \|u\| \leqslant n^{-1/2}} |K(y) - K(y+u)|.$$

Then, for any $x \in \{\|x\| \leqslant n^r\}$, choosing k such that $\|x - x_k\| \leqslant \varepsilon_n$, simple algebra gives

$$\begin{aligned}
|Z_n(x)| &\leqslant |Z_n(x_k)| + \sum_{i=1}^n \frac{|f(X_i)(K_{h_i}(x-X_i) - K_{h_i}(x_k - X_i))|}{q_{i-1}(X_i)} \\
&\quad + \sum_{i=1}^n \int |f(y)(K_{h_i}(x-y) - K_{h_i}(x_k-y))|\,\mathrm{d}y \\
&= |Z_n(x_k)| + \sum_{i=1}^n \frac{f(X_i) h_i^{-d} r_n((x_k - X_i)/h_i)}{q_{i-1}(X_i)} + \sum_{i=1}^n h_i^{-d} \int f(y) r_n((x_k-y)/h_i)\,\mathrm{d}y \\
&\leqslant |Z_n(x_k)| + |Z_{n,2}(x_k)| + 2\sum_{i=1}^n h_i^{-d} \int f(y) r_n((x_k-y)/h_i)\,\mathrm{d}y
\end{aligned}$$

where

$$Z_{n,2}(x) = \sum_{i=1}^n \left\{ \frac{f(X_i)}{q_{i-1}(X_i)} h_i^{-d} r_n((x_k - X_i)/h_i) - \int f(y) h_i^{-d} r_n((x_k-y)/h_i)\,\mathrm{d}y \right\}.$$

It follows that

$$\begin{aligned}
|Z_n(x)| &\leqslant |Z_n(x_k)| + |Z_{n,2}(x_k)| + 2U_f \sum_{i=1}^n \int r_n(u)\,\mathrm{d}u \\
&\leqslant \max_{k=1,\ldots N} \big\{|Z_n(x_k)| + |Z_{n,2}(x_k)|\big\} + 2U_f C_K n^{1/2}.
\end{aligned} \tag{10}$$

The second term of the previous bound being negligible compared to the target bound $\sqrt{n\log(n)/\lambda_n h_n^d}$, it remains to bound the first term. This is done by applying Theorem 3 with

$$Y_i = \frac{f(X_i)}{q_{i-1}(X_i)} K_{h_i}(x_k - X_i)^2 - \int f(x) K_{h_i}(x_k - x) dx.$$

Using that $q_{i-1} \geqslant \lambda_{i-1} q_0$, we get by (H1),

$$\begin{aligned}
\int \frac{f(y)^2 K_{h_i}(x_k-y)^2}{q_{i-1}(y)}\,\mathrm{d}y &\leqslant h_i^{-d} \int \frac{f(y)^2 K((x_k-y)/h_i)^2}{\lambda_{i-1} q_0(y)} \frac{\mathrm{d}y}{h_i^d} \\
&\leqslant \lambda_{i-1}^{-1} h_i^{-d} U_f U_{f,q_0} v_K,
\end{aligned}$$

with $v_K = \int K(u)^2\,\mathrm{d}u < \infty$. We have (bound on the quadratic variation)

$$\begin{aligned}
v &= \sum_{i=1}^n E\left[\left(f(X_i)\frac{K_{h_i}(x_k - X_i)}{q_{i-1}(X_i)} - \int f(y)K_{h_i}(x_k - y)\right)^2 \mid \mathscr{F}_{i-1}\right] \\
&\leqslant \sum_{i=1}^n \int \frac{f(y)^2 K_{h_i}(x_k - y)^2}{q_{i-1}(y)}\,\mathrm{d}y \\
&\leqslant U_f U_{f,q_0} v_K \sum_{i=1}^n \lambda_{i-1}^{-1} h_i^{-d},
\end{aligned}$$

and using that λ_n and h_n are decreasing, we obtain that

$$v \leqslant U_{f,q_0} v_K n \lambda_n^{-1} h_n^{-d}.$$

Similarly (uniform bound on martingale increments)

$$\begin{aligned}
m &= \max_{i=1,\ldots,n} \sup_{y\in\mathbb{R}^d} \left| \frac{f(y) h_i^{-d} K((x_k - y)/h_i)}{q_{i-1}(y)} - \int f(x) K_{h_i}(x_k - x)dx \right| \\
&\leqslant U_{f,q_0} U_K \lambda_n^{-1} h_n^{-d} + U_f.
\end{aligned}$$

We are in position to apply the concentration inequality for martingales given in Theorem 3, leading to, for any $t \geqslant 0$

$$\begin{aligned}
\mathbb{P}\left(\max_{k=1,\ldots N} |Z_n(x_k)| > t\right) &\leqslant N \max_{k=1,\ldots N} \mathbb{P}\left(|Z_n(x_k)| > t\right) \\
&\leqslant 2N \max_{k=1,\ldots N} \exp\left(-\frac{t^2}{2(v + tm/3)}\right) \\
&= 2N \exp\left(-\frac{Ct^2 \lambda_n h_n^d}{n+t}\right)
\end{aligned}$$

for some constant C depending only on (f, q_0, K). Very similar algebra with $\max_{k=1,\ldots N} |Z_{n,2}(x_k)|$ leads now to a quadratic variation bounded by

$$v_2 \leqslant U_f U_{f,q_0} U_K C_K n h_n^{-d} \lambda_n^{-1} n^{-1/2}$$

(this is obtained using (H2) : $\int r_n^2 \leqslant U_K \int r_n \leqslant U_K C_K n^{-1/2}$) and

$$m_2 \leqslant U_{f,q_0} U_K \lambda_n^{-1} h_n^{-d}$$

Since $v_2 = O(v)$ and $m_2 = O(m)$, a bound concerning Z_n will be valid for $Z_{n,2}$.

Hence we focus on Z_n in the next few lines. From the assumption that $\log(n)/nh_n^d = O(\lambda_n)$, we deduce that $h_n^{-1} \ll n^{1/d}$, thus we have $N \propto \varepsilon_n^{-d} n^{rd} \ll n^{d/2+d+rd}$. We can conclude by choosing $t = \gamma\sqrt{n\log(n)/h_n^d\lambda_n}$ with γ large enough, we get (because $\log(n)/(nh_n^d) = O(\lambda_n)$, $t = O(n)$))

$$\sum_{n\geqslant 1} \mathbb{P}\left(\max_{k=1,\ldots N} |Z_n(x_k)| > \gamma\sqrt{n\log(n)/h_n^d\lambda_n}\right) < +\infty,$$

which in light of the Borel-Cantelli lemma implies that, almost surely, $\max_{k=1,\ldots N} |Z_n(x_k)| = O(\sqrt{n\log(n)/h_n^d\lambda_n})$. $\square$

A.4 Proof of Theorem 6

The proof is similar to the proof of Lemma 5, based on (8), except that the bound in $a_n\lambda_n^{-1/2}$ on the variance term $f_n - \widetilde{f}_n$, will be improved in Lemma 12. Our ability to improve the initial bound $a_n\lambda_n^{-1/2}$ follows from this useful technical lemma in which some bounds are established on f^k/q_i, $k = 1, 2$.

Define the family of events, for any $r > 0$ and $U > 0$,

$$E_{U,r} = \left\{ \omega : \sup_{n\geqslant 1} \left\{ \frac{\sup_{\|x\|\leqslant n^r} |f_n(x) - f(x)|}{a_n\lambda_n^{-1/2} + \overline{h_n^2}} \right\} \leqslant U \right\}.$$

Lemma 11. *Under (H1), (H2), (H3), (H4), if λ_n and h_n are positive decreasing sequences, for any $r > 0$,*

$$\sup_{\|x\|\geqslant n^r} f(x) \ll a_n\lambda_n^{-1/2} + \overline{h_n^2}. \tag{11}$$

If in addition there exists $\delta > 0$ for which $a_n^{1-\delta} \ll \lambda_n$, $\overline{h_n^2}^{1-\delta} \ll \lambda_n$ and λ_n is not the constant sequence equal to 1. Then for any $U > 0$, $r > 0$, there exists $A > 0$ such that under $E_{U,r}$, we have for all $n \geqslant 1$ and $x \in \mathbb{R}^d$,

$$\frac{f(x)}{q_n(x)} \leqslant A a_n^{-1}, \tag{12}$$

$$\frac{f(x)^2}{q_n(x)} \leqslant A q_0(x). \tag{13}$$

Proof. Let $c_n = a_n\lambda_n^{-1/2} + \overline{h_n^2}$. Under (H4), for $x \geqslant n^r$, $\varepsilon > 0$, $f(x)^\varepsilon \leqslant U_{f^\varepsilon,q_0} q_0(x) \leqslant U_{f^\varepsilon,q_0} C_0 n^{-rk_0}$. It suffices now to choose ε small enough such that $n^{-rk_0} \ll c_n^\varepsilon$. This is made possible because, h_n and λ_n being decreasing,

$$c_n \geqslant \sqrt{\frac{\log n}{nh_n^d}}\lambda_n^{-1/2} + h_n^2 \geqslant \sqrt{\frac{\log n}{nh_0^d}}\lambda_0^{-1/2} + h_n^2$$

and $nh_n^d \to \infty$ since a_n is bounded.

Let $U > 0$, $r > 0$. From what has been established just before, there exists N_0 such that for all $n \geqslant N_0$, $f(x) \geqslant 2Uc_n$ implies that $\|x\| \leqslant n^r$. Because $q_n(x) \geqslant (1 - \lambda_n) f_n(x)$, under $E_{U,r}$, for all $n \geqslant N_0$ and $f(x) \geqslant 2Uc_n$,

$$\begin{aligned}
\frac{f(x)}{q_n(x)} &\leqslant (1-\lambda_n)^{-1}\left(1 + \frac{f(x) - f_n(x)}{f_n(x)}\right) \\
&\leqslant (1-\lambda_n)^{-1}\left(1 + \frac{|f(x) - f_n(x)|}{f(x) - |f(x) - f_n(x)|}\right) \\
&\leqslant 2(1-\lambda_n)^{-1}.
\end{aligned}$$

Furthermore, for any $1 \leqslant n \leqslant N_0$, and $x \in \mathbb{R}^d$, $f(x)/q_n(x) \leqslant U_{f,q_0}/\lambda_{N_0}$. This implies that under $E_{U,r}$, for all $n \geqslant 1$ and $f(x) \geqslant 2Uc_n$,

$$\frac{f(x)}{q_n(x)} \leqslant M,$$

where M depends on U and the sequences of interest. As a consequence, for all $n \geqslant 1$ and $x \in \mathbb{R}^d$,

$$\frac{f(x)}{q_n(x)} \leqslant M + \frac{(2Uc_n)^{1-\delta}}{\lambda_n}\frac{f(x)^\delta}{q_0(x)},$$
$$\frac{f(x)^2}{q_n(x)} \leqslant Mf(x) + \frac{(2Uc_n)^{2-\delta}}{\lambda_n}\frac{f(x)^\delta}{q_0(x)} \leqslant \left(MU_{f,q_0} + \frac{(2Uc_n)^{2-\delta}}{\lambda_n}U^2_{f^{\delta/2},q_0}\right)q_0(x).$$

Using that $(a+b)^\delta \leqslant 2^\delta(a^\delta + b^\delta)$ and by assumption on the sequences h_n, λ_n,

$$\frac{c_n^{1-\delta}}{\lambda_n} \leqslant 2\left(\frac{a_n^{1-\delta}}{\lambda_n^{(3-\delta)/2}} + \frac{\overline{h_n^2}^{1-\delta}}{\lambda_n}\right) = O\left(a_n^{1-\delta-(1-\delta)(3-\delta)/2} + 1\right) = O\left(a_n^{-1} + 1\right),$$
$$\frac{c_n^{2-\delta}}{\lambda_n} \leqslant 4\left(\frac{a_n^{2-\delta}}{\lambda_n^{(4-\delta)/2}} + \frac{\overline{h_n^2}^{2-\delta}}{\lambda_n}\right) \to 0. \tag{14}$$

By taking A large enough, we get the statement. □

Lemma 12. *Under (H1), (H2), (H3), (H4), if λ_n and h_n are positive decreasing sequences such that there exists $\delta > 0$ for which $a_n^{1-\delta} \ll \lambda_n$, $\overline{h_n^2}^{1-\delta} \ll \lambda_n$, we have,*

$$\sup_{\|x\|\leqslant n^r} |Z_n(x)| = O_{\mathbb{P}}\left(\sqrt{\frac{n\log(n)}{h_n^d}}\right)$$
$$M_n = O_{\mathbb{P}}\left(\sqrt{n}\right).$$

As a consequence, $\sup_{\|x\|\leqslant n^r} |f_n(y) - \widetilde{f}_n(y)| = O_{\mathbb{P}}(a_n)$.

Proof. Note that because $f_n(y) - \widetilde{f}_n(y) = (Z_n(y) - \widetilde{f}_n(y)M_n)/(M_n + n)$, the final result is a consequence the bounds given for Z_n and M_n.

When λ_n is the constant sequence equal to 1 the bounds on Z_n and M_n are already given by Lemma 10. Consider the case where λ_n is not the constant sequence equal to 1. We start doing the same as in the proof of Lemma 5 to get (10). Since the second term has the stated rate of convergence, we only have to show that for any $\epsilon > 0$, there exists $\gamma > 0$ such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\max_{k=1,\ldots N}\left\{|Z_n(x_k)| + U_f|Z_{n,2}(x_k)|\right\} > \gamma\sqrt{n\log(n)/h_n^d}\right) \leqslant \varepsilon. \tag{15}$$

Let $\epsilon > 0$. By virtue of the monotone convergence theorem, the exists $U > 0$, large enough, such that $\mathbb{P}(E^c_{U,r}) \leqslant \epsilon$. On $E_{U,r}$, using (13), the quadratic variation is bounded by

$$\sum_{i=1}^n \int \frac{f(y)^2}{q_{i-1}(y)} K_{h_i}(x_k - y)^2 \,\mathrm{d}y \leqslant A\|q_0\|_\infty v_K n h_n^{-d},$$

and, based on (12), a uniform bound on martingale increments is

$$\begin{aligned} m &= \max_{i=1,\ldots,n}\sup_{y\in\mathbb{R}^d}\left|\frac{f(y)h_i^{-d}K((x_k-y)/h_i)}{q_{i-1}(y)} - \int f(x)K_{h_i}(x_k - x)dx\right| \\ &\leqslant \max_{i=1,\ldots,n} AU_K a_i^{-1}h_i^{-d} + U_f \\ &= \max_{i=1,\ldots,n} AU_K \log(i)^{-1/2} i^{1/2} h_i^{-d/2} + U_f \\ &\leqslant AU_K\sqrt{\frac{nh_n^{-d}}{\log(n)}} + U_f. \end{aligned}$$

Theorem 3 implies that for any $t > 0$

$$\mathbb{P}\left(\max_{k=1,\dots N} |Z_n(x_k)| > t, E_{U,r}\right) = 2N \exp\left(-\frac{Ct^2 h_n^d}{n + t\sqrt{nh_n^d/\log(n)}}\right)$$

where C depends on (A, f, K) only. Very similar algebra with $\max_{k=1,\dots N} |Z_{n,2}(x_k)|$ gives

$$\mathbb{P}\left(\max_{k=1,\dots N} |Z_{n,2}(x_k)| > t, E_{U,r}\right) \leqslant 2N \exp\left(-\frac{C' t^2 h_n^d}{n\lambda_n + t\sqrt{nh_n^d/\log(n)}}\right).$$

where C' depends on (A, f, K) only. Choosing $t = \gamma\sqrt{n\log(n)/h_n^d}$, with γ large enough, both previous bounds go to 0. This combined with $\mathbb{P}(E_{U,r}^c) \leqslant \epsilon$ leads to (15).

The bounds on the quadratic variation and the increments of M_n is derived in a very similar way as before with, by (12) and (13),

$$\sum_{i=1}^n \int \frac{f(y)^2}{q_{i-1}(y)} \, \mathrm{d}y \leqslant nA = v$$
$$\max_{i=1,\dots,n} \sup_{y\in\mathbb{R}^d} \left| \frac{f(y)}{q_{i-1}(y)} - 1 \right| \leqslant \max_{i=1,\dots,n} Aa_i^{-1} + 1 \leqslant A\sqrt{n} + 1 = m.$$

Theorem 3 implies that for any $t > 0$

$$\mathbb{P}\left(|M_n| > t, E_{U,r}\right) = \exp\left(-\frac{Ct^2}{n + t\sqrt{n}}\right).$$

Choosing $t = \gamma\sqrt{n}$ with γ large gives the result.

□

A.5 Proof of Theorem 7

Because $\log(n)/n \ll \lambda_n$, applying Lemma 10 gives that $I_n(1)$ converges in probability to 1. Because of (4), the problem reduces to the estimation of the limit of $n^{1/2}((I_n(g) - \int gf), (I_n(1) - 1))$. Verifying the conditions of Theorem 4 with

$$W_{n,i} = \frac{1}{\sqrt{n}} \Big(\frac{g(X_i)f(X_i)}{q_{i-1}(X_i)} - \int gf \Big),$$

will prove the convergence of $n^{1/2}(I_n(g) - \int gf)$ to the limit $\mathcal{N}(0, V(f, g)$. This will imply that $n^{1/2}(I_n(1) - 1)$ converges to zero in probability since $V(f, 1) = 0$, and the result will follow by virtue of Slutsky's Lemma.

Equation (5) is satisfied. We now show (6) with $v_* = V(f, g) = \int g^2 f - (\int fg)^2$, or equivalently that

$$n^{-1} \sum_{i=1}^n \int \frac{g^2 f^2}{q_{i-1}} \to \int g^2 f, \qquad \text{in probability.}$$

We will prove that

$$n^{-1} \sum_{i=1}^n \int \Big| \frac{g^2 f^2}{q_{i-1}} - g^2 f \Big|, \qquad \text{in probability.} \tag{16}$$

From Lemma 5, with probability 1, there exists $0 < U < +\infty$ such that

$$\sup_{n\geqslant 1}\sup_{\|x\|\leqslant n^r}\frac{|q_n(x)-f(x)|}{\widetilde{c}_n} < U,$$

with $\widetilde{c}_n = a_n\lambda_n^{-1/2} + \overline{h_n^2} + \lambda_n$. This implies that, for any $\gamma > 0$, if $f(x) > (1+\gamma)U\widetilde{c}_n$, then, by (11) $\|x\| \leqslant n^r$ (for all n if U is large enough), and

$$\frac{|q_n(x)-f(x)|}{q_n(x)} \leqslant \frac{|q_n(x)-f(x)|}{(1+\gamma)U\widetilde{c}_n - |q_n(x)-f(x)|} < \frac{1}{\gamma}.$$

Splitting each integral in (16) gives that, with probability 1,

$$\begin{aligned}
&\int\left|\frac{g^2f^2}{q_{i-1}} - g^2f\right| \\
&\leqslant \int_{f>(1+\gamma)U\widetilde{c}_i}\left\{g^2f\frac{|q_{i-1}-f|}{q_{i-1}}\right\} + \lambda_i^{-1}\int_{f\leqslant(1+\gamma)U\widetilde{c}_i}\left\{\frac{g^2f^2}{q_0}\right\} + \int_{f\leqslant(1+\gamma)U\widetilde{c}_i} g^2f \\
&\leqslant \frac{1}{\gamma}\int g^2f + ((1+\gamma)U\widetilde{c}_i)^{2-\delta}\lambda_i^{-1}\int\frac{g^2f^\delta}{q_0} + ((1+\gamma)U\widetilde{c}_i)^{1-\delta}\int g^2f^\delta.
\end{aligned}$$

Using that, from (H4) and (H5), $\int g^2f^\delta/q_0 = \int(g^2f^\delta/{q_0}^{k+1})q_0^k \leqslant C\int g^2{q_0}^k < \infty$ for some $C > 0$, and invoking (14), the term in the middle goes to 0. Similarly, the right-hand side term goes to 0. We obtain, invoking Cesaro's lemma, that, with probability 1,

$$\limsup_{n\to\infty}\left|n^{-1}\sum_{i=1}^n\int\left\{\frac{g^2f^2}{q_{i-1}} - g^2f\right\}\mathrm{d}x\right| \leqslant \frac{1}{\gamma}\int g^2f. \tag{17}$$

But γ is arbitrary.

Finally, we verify the Lindeberg condition (7). We have to prove that

$$n^{-1}\sum_{i=1}^n\int\frac{g(x)^2f(x)^2}{q_{i-1}(x)}\mathrm{I}_{\{|g(x)f(x)/q_{i-1}(x)-I_g|>\varepsilon\sqrt{n}\}}\,\mathrm{d}x \to 0 \quad \text{in probability.}$$

Set $A_{i,n} = \{|g(x)f(x)/q_{i-1}(x) - I_g| > \varepsilon\sqrt{n}\}$, from (17) it follows that

$$\begin{aligned}
n^{-1}\sum_{i=1}^n\int\frac{g(x)^2f(x)^2}{q_{i-1}(x)}\mathrm{I}_{A_{i,n}}(x)\,\mathrm{d}x &= o(1) + n^{-1}\sum_{i=1}^n\int g(x)^2f(x)\mathrm{I}_{A_{i,n}}(x)\,\mathrm{d}x \\
&= o(1) + \int g(x)^2f(x)\Big(n^{-1}\sum_{i=1}^n\mathrm{I}_{A_{i,n}}(x)\Big)\,\mathrm{d}x.
\end{aligned}$$

Consider n such that $I_g \leqslant \varepsilon n/2$, then $A_{i,n} \subset \{g(x)f(x)/q_{i-1}(x) > \varepsilon\sqrt{n}/2\}$. Since $q_{i-1}(x) \geqslant \lambda_i q_0(x)$ and $\lambda_i \gg a_i^{1-\delta} \geqslant ci^{-(1-\delta)/2} \geqslant cn^{-(1-\delta)/2}$, we obtain

$$A_{i,n} \subset B_n = \{g(x)f(x)/q_0(x) > c\varepsilon n^{\delta/2}/2\}.$$

Finally

$$n^{-1}\sum_{i=1}^n\int\frac{g(x)^2f(x)^2}{q_{i-1}(x)}\mathrm{I}_{A_{i,n}}(x)\,\mathrm{d}x \leqslant o(1) + \int g(x)^2f(x)I_{B_n}(x)\,\mathrm{d}x$$

and we conclude with the Lebesgue theorem. □

Appendix B The compact case

In this section we present a bound for the variance term $f_n - \widetilde{f}_n$, which is analogous to the one of Theorem 12. The bound on the bias term $\widetilde{f}_n - f$ can be easily treated analogously to Lemma 9, under suitable assumptions.

(H6) The support of f, S_f, is compact and for all $x \in S_f$ we have $L_f \leqslant f(x) \leqslant U_f$. For all $x \in S_f$, $q_0(x) \geqslant L_{q_0} > 0$.

In addition

$$\min_{x \in S_f} \min_{h \leqslant h_1} (\mathrm{I}_{\{S_f\}} * K_h)(x) = C_{SK} > 0. \tag{18}$$

It is not difficult to prove that (18) is satisfied if S_f is convex (since it is also bounded). The following event

$$E_L = \{\omega \,:\, \forall n \geqslant 1 \ \inf_{y \in S_f} q_n(y) \geqslant L\},$$

will play an important role in the following. We state the key property related to E_L in the following lemma.

Lemma 13. *Under (H1), (H2), (H6), if λ_n and h_n are positive decreasing sequences such that $\log(n)/nh_n^d \ll \lambda_n$, then with probability* 1,

$$\liminf_{n \to \infty} \inf_{x \in S_f} q_n(x) \geqslant L_f C_{SK}.$$

Moreover $\mathbb{P}(E_L) \to 1$ as $L \to 0$.

Proof. Recall (8) and apply Lemma 10 (using that $\log(n)/nh_n^d \ll \lambda_n$ implies that $\log(n)/n \ll \lambda_n$) to obtain that $M_n = o(n)$. Finally, applying again Lemma 10, and identity (9), we obtain that with probability 1,

$$\sup_{y \in S_f} |f_n(y) - \widetilde{f}_n(y)| \to 0.$$

Now write,

$$q_n(y) \geqslant (1 - \lambda_n) f_n \geqslant (1 - \lambda_n)(\widetilde{f}_n(y) - |f_n(y) - \widetilde{f}_n(y)|).$$

Using (18) gives that

$$\inf_{y \in S_f} q_n(y) \geqslant (1 - \lambda_n) L_f C_{SK} - \sup_{y \in S_f} |f_n(y) - \widetilde{f}_n(y)|.$$

Taking the limit permits to obtain the first statement. By assumption, the variable $U_n = \inf_{y \in S_f} q_n(y)$ satisfies $U_n \geqslant \lambda_n L_{q_0}$; since in addition $\liminf_n U_n \geqslant L_f C_{SK}$, we have $\inf_n U_n > 0$ w.p.1. Hence $\mathbb{P}(E_L) \to 1$ as $L \to 0$. □

This result shows that the conditions on λ_n are weakened in the compact case as $a_n^{2-\delta} \ll \lambda_n$ is replaced by $a_n^2 \ll \lambda_n$.

Theorem 14 (compact case)**.** *Under (H1), (H2), (H6), if λ_n and h_n are positive decreasing sequences such that $a_n^2 \ll \lambda_n$, we have*

$$\begin{aligned}
&\sup_{\|x\|\leqslant n^r} |Z_n(x)| = O_{\mathbb{P}}\left(\sqrt{\frac{n\log(n)}{h_n^d}}\right)\\
&M_n = O_{\mathbb{P}}(\sqrt{n}).
\end{aligned}$$

As a consequence,

$$\sup_{\|x\|\leqslant n^r} |f_n(y) - \widetilde{f}_n(y)| = O_{\mathbb{P}}\left(\sqrt{\frac{\log(n)}{nh_n^d}}\right).$$

Proof. We start doing the same as in the proof of Lemma 10 and getting (10). Since the second term has the stated rate of convergence, it remains to show that $\max_{k=1,\dots N}\{|Z_n(x_k)| + U_f|Z_{n,2}(x_k)|\}$ has the right order.

Let $\epsilon > 0$. We can split according to two events E_L and E_L^c with $L > 0$ small enough such that, by virtue of Lemma 13, for all $t \geqslant 0$,

$$\begin{aligned}
&\mathbb{P}\Big(\max_{k=1,\dots N}\{|Z_n(x_k)| + U_f|Z_{n,2}(x_k)|\} > t\Big)\\
&\qquad\qquad \leqslant \mathbb{P}\Big(\max_{k=1,\dots N}\{|Z_n(x_k)| + U_f|Z_{n,2}(x_k)|\} > t, E_L\Big) + \epsilon.
\end{aligned}$$

We will now apply Theorem 3 to the martingale $Z_n(x_k)$. The quadratic variation is

$$\begin{aligned}
v &= \sum_{i=1}^n E\left[\left(\frac{f(X_i)}{q_{i-1}(X_i)}K_{h_i}(x_k - X_i) - \int f(y)K_{h_i}(x_k-y)\,\mathrm{d}y\right)^2 \Big| \mathscr{F}_{i-1}\right]\\
&\leqslant \sum_{i=1}^n \int_{S_f} \frac{f(y)^2}{q_{i-1}(y)}K_{h_i}(x_k-y)^2\,\mathrm{d}y\\
&= U_f^2 L^{-1} v_K \sum_{i=1}^n h_i^{-d}\\
&= U_f^2 L^{-1} v_K n h_n^{-d}.
\end{aligned}$$

A uniform bound on martingale increments is

$$m = \max_{i=1,\dots,n}\sup_{y\in S_f}\left|\frac{f(y)h_i^{-d}K((x_k-y)/h_i)}{q_{i-1}(y)} - \int f(x)K_{h_i}(x_k-x)dx\right| \leqslant U_f U_K L^{-1} h_n^{-d} + U_f.$$

As in the proof of Lemma 5, Theorem 3 implies that for any $t > 0$

$$\mathbb{P}\left(\max_{k=1,\dots N}|Z_n(x_k)| > t, E_b\right) = 2N\exp\left(-\frac{Ct^2h_n^d}{n+t}\right)$$

where C depends on (L, f, K) only. Very similar algebra with $\max_{k=1,\dots N}|Z_{n,2}(x_k)|$ gives

$$\mathbb{P}\left(\max_{k=1,\dots N}|Z_{n,2}(x_k)| > t\right) \leqslant 2N\exp\left(-\frac{C't^2h_n^d}{n\lambda_n + t}\right).$$

where C' depends on (L, f, K) only. Choosing $t = \gamma\sqrt{n\log(n)/h_n^d}$, with γ large enough, the both previous bounds go to 0. Hence we get

$$\limsup_{n\to\infty} \mathbb{P}\left(\max_{k=1,\dots N}\{|Z_n(x_k)| + U_f|Z_{n,2}(x_k)|\} > \gamma\sqrt{n\log(n)/h_n^d}\right) \leqslant \varepsilon,$$

which is the statement of the theorem.

For the second statement, we will use (9). Let us apply Theorem 3 with

$$Y_i = \frac{f(X_i)}{q_{i-1}(X_i)} - \int f(x)dx.$$

On the set E_L, we have (bound on the quadratic variation)

$$v = \sum_{i=1}^n E\left[\left(\frac{f(X_i)}{q_{i-1}(X_i)} - \int f(x)dx\right)^2 \Big|\mathscr{F}_{i-1}\right] \leqslant \sum_{i=1}^n \int \frac{f(y)^2}{q_{i-1}(y)}\,\mathrm{d}y \leqslant U_f L^{-1} n.$$

Still on E_L, a bound on the martingale increments is given as

$$m = \max_{i=1,\dots,n} \sup_{y\in\mathbb{R}^d} \left|\frac{f(y)}{q_{i-1}(y)} - 1\right| \leqslant U_f L^{-1} + 1.$$

Theorem 3 implies that for any $t > 0$

$$\mathbb{P}\left(|M_n| > t, E_L\right) \leqslant 2\exp\left(-\frac{Ct^2}{n+t}\right).$$

for some $C > 0$ depending only on (f, L). Choosing $t = \gamma\sqrt{n}$ with $\gamma > 0$ large enough and $L > 0$ small enough, we get that

$$\limsup_{n\to\infty} \mathbb{P}\left(|M_n| > \gamma\sqrt{n}\right) \leqslant \epsilon.$$

Using (8), we directly get the third statement. □

References

Azaïs, R., B. Delyon, and F. Portier (2018). Integral estimation based on markovian design. *Advances in Applied Probability 50*(3), 833–857.

Bertail, P. and F. Portier (2018). Rademacher complexity for markov chains: Applications to kernel smoothing and metropolis-hasting. *To appear in Bernoulli*.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities*. Oxford University Press.

Cappé, O., R. Douc, A. Guillin, J.-M. Marin, and C. P. Robert (2008). Adaptive importance sampling in general mixture classes. *Statistics and Computing 18*(4), 447–459.

Cappé, O., A. Guillin, J.-M. Marin, and C. P. Robert (2004). Population monte carlo. *Journal of Computational and Graphical Statistics 13*(4), 907–929.

Chopin, N. (2004). Central limit theorem for sequential monte carlo methods and its application to bayesian inference. *The Annals of Statistics 32*(6), 2385–2411.

Del Moral, P., A. Doucet, and A. Jasra (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(3), 411–436.

Delyon, B. and F. Portier (2018). Asymptotic optimality of adaptive importance sampling. *Thirty-second Conference on Neural Information Processing Systems*.

Douc, R., A. Guillin, J.-M. Marin, and C. P. Robert (2007a). Convergence of adaptive mixtures of importance sampling schemes. *The Annals of Statistics*, 420–448.

Douc, R., A. Guillin, J.-M. Marin, and C. P. Robert (2007b). Minimum variance importance sampling via population monte carlo. *ESAIM: Probability and Statistics 11*, 427–447.

Elvira, V., L. Martino, D. Luengo, and M. F. Bugallo (2015). Generalized multiple importance sampling. *arXiv preprint arXiv:1511.03095*.

Evans, M. and T. Swartz (2000). *Approximating integrals via Monte Carlo and deterministic methods*. Oxford Statistical Science Series. Oxford University Press, Oxford.

Feng, M. B., A. Maggiar, J. Staum, and A. Wächter (2018). Uniform convergence of sample average approximation with adaptive multiple importance sampling. In *2018 Winter Simulation Conference (WSC)*, pp. 1646–1657. IEEE.

Freedman, D. A. (1975). On tail probabilities for martingales. *Ann. Probability 3*, 100–118.

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, 1317–1339.

Giné, E. and A. Guillou (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincaré Probab. Statist. 37*(4), 503–522.

Giné, E. and A. Guillou (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist. 38*(6), 907–921. En l'honneur de J. Bretagnolle, D. Dacunha-Castelle, I. Ibragimov.

Givens, G. H. and A. E. Raftery (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *Journal of the American Statistical Association 91*(433), 132–141.

Hall, P. and C. C. Heyde (1980). *Martingale limit theory and its application*. Academic Press.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory 24*(3), 726–748.

Kloek, T. and H. K. Van Dijk (1978). Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, 1–19.

Marin, J.-M., P. Pudlo, and M. Sedki (2019). Consistency of adaptive importance sampling and recycling schemes. *To appear in Bernoulli*.

Neddermeyer, J. C. (2009). Computationally efficient nonparametric importance sampling. *Journal of the American Statistical Association 104*(486), 788–802.

Oh, M.-S. and J. O. Berger (1992). Adaptive importance sampling in Monte Carlo integration. *J. Statist. Comput. Simulation 41*(3-4), 143–168.

Owen, A. and Y. Zhou (2000). Safe and effective importance sampling. *J. Amer. Statist. Assoc. 95*(449), 135–143.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples*.

Owen, A. B. and Y. Zhou (2019). The square root rule for adaptive importance sampling. *arXiv preprint arXiv:1901.02976*.

Pollard, D. (1984). *Convergence of stochastic processes*. Springer.

Robert, C. P. and G. Casella (2004). *Monte Carlo statistical methods* (Second ed.). Springer Texts in Statistics. Springer-Verlag, New York.

van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

West, M. (1993). Approximating posterior distributions by mixtures. *Journal of the Royal Statistical Society: Series B (Methodological) 55*(2), 409–422.

Zhang, P. (1996). Nonparametric importance sampling. *J. Amer. Statist. Assoc. 91*(435), 1245–1253.